DOCUMENT RESUME

ED 374 572 EA 026 159

AUTHOR Banks, Karen E.
TITLE Assessment's Conflicting Purposes, Conflicting Politics: Impact on Local School Systems.
PUB DATE Apr 94
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability; Educational Assessment; *Educational Policy; Elementary Secondary Education; Local Issues; *Political Influences; Public Schools; *School Districts; *Standardized Tests; State School District Relationship; *State Standards
IDENTIFIERS Wake County Public School System NC

ABSTRACT

This paper examines political influences on assessment programs and their effects on local school systems. Specifically, it describes North Carolina's political climate and examines the impact of political influences on educational assessment in the state. The political climate in North Carolina has produced a plethora of state-mandated tests, some of which include the End-of-Grade (EOG) Test; Writing Assessment; and the High School End-of-Grade Test in algebra or geometry. Problems with the testing system are described in detail. Clearly, the purpose of local assessments--to improve instruction--may differ from the purposes of externally mandated assessments. To reduce conflicts with state policy makers, local educators are advised to: (1) maintain their own professional integrity; (2) organize to voice concerns; (3) make noise; and (4) deemphasize the importance of test scores. State officials can help by including more teachers and local education staff in developing report formats and test frameworks and ensuring that tests accurately measure expectations for student performance. Ideally, future state and national assessments will be based on what is best for students and teachers. Six exhibits are included. The appendix contains a review of research that explored the effects of accountability efforts on instruction. (LMI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

94.03A

ED 374 572

Assessment's Conflicting Purposes, Conflicting Politics: Impact on Local School Systems

Karen E. Banks
Wake County Public School System
Raleigh, N.C.

Paper Presented to the American Educational Research Association
April, 1994

The opinions expressed in this paper reflect those of the author and not necessarily those of the Wake County Public School System.

EA 026 159

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. Banks

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

94.03A

Assessment's Conflicting Purposes, Conflicting Politics: Impact on Local School Systems

Karen E. Banks
Wake County Public School System
Raleigh, N.C.

Introduction

This paper examines political influences on assessment programs and how they affect local school systems. Reviewing the purposes of assessment may be a good place to start, because it is difficult to recognize distortion unless one begins with a clear picture. In November, 1993, a subcommittee of the National Goals Panel published the report, "*Promises to Keep: Creating High Standards for American Students.*" In this report, the committee described several purposes for an assessment system:

> "A system of student assessments linked to world-class standards would provide information that could be used to:
>
> - exemplify for students, parents, and teachers the kinds and levels of achievement expected;
> - improve classroom instruction and learning outcomes for all students;
> - inform students, parents, and teachers about student progress;
> - measure and hold students, schools, school districts, states, and the Nation accountable for educational performance; and
> - assist education policymakers with programmatic decisions.
>
> It is unlikely that all of these purposes could be accomplished with the same assessment."

At first glance, most educators would probably embrace this description because it presents several positive features. For example, the description recognizes the importance of increased learning outcomes for all students. In addition, the description acknowledges the need for multiple measures or tests. Combining the use of multiple measures with sampling will allow less testing of each individual student, because no single test will have to address all purposes.

Closer examination of this description, however, reveals some troubling issues. The problems with the description illustrate one important theme of this paper. The description fails to explicitly infuse improvement of student learning into each purpose. "Exemplify kinds and levels of expected achievement" is listed before "improving instruction and learning." *I believe the* ***primary*** *function of tests is to improve instruction and learning. More broadly, we reform not for the sake of reform, but to improve instruction and learning.*

Challenges of Politics, Reform, and Accountability

Educators, test developers, and policymakers have long struggled with the challenge of designing assessment systems that fulfill multiple purposes. The challenge includes balancing costs, technical merit, instructional utility, and instructional time lost to testing. The challenge has grown more difficult because of current emphases on state-mandated tests as tools of both **reform** and **accountability.** Further, these recent emphases have moved assessment into the political arena to an unprecedented extent and increased the conflicts between educators and policymakers who mandate assessments. Reform and accountability appear to be on a collision course, with local school systems caught in the middle.

Political factors affecting assessment programs include public opinion, partisan disputes, campaign rhetoric, and legislative micromanagement of assessment programs. These factors can affect the design, development, selection, and reporting of assessments, as well as the kinds of activities local school districts undertake in response to the assessments. Overall, these political factors also affect the impact of assessments on accountability and reform efforts.

Some knowledgeable experts believe that assessments can facilitate or even drive *reforms*, at least under certain circumstances (Popham, 1993; Corbett and Wilson, 1991). Others contend assessments cannot effectively drive reforms (Cuban, 1993). In real-world settings, the efficacy of mandating reforms by mandating assessments remains unclear.

Assessment as a tool for *accountability*, on the other hand, remains a much easier goal to accomplish, but carries potentially more damaging outcomes. Although testing for accountability is not new, the current extent of such testing and public scrutiny of the results *are* unprecedented. In many cases, mandated testing for accountability has led to narrowing of the curriculum, questionable methods of raising test scores, and a "quick fix" approach to reform (see Appendix 1).[1] Some would argue we need to reconsider the viability of such assessments to accurately measure achievement, at least when the stakes are high (Anrig, 1991; Bond, Friedman, and vander Ploeg, 1993; Stiggins, 1993). In summary, no consensus exists about what assessment can and cannot accomplish.

---

[1]Ironically, the most frequently proposed method of alleviating these problems is to develop newer, more authentic assessments of student achievement. One wonders why so many individuals are so enamored with tests when many previous efforts have failed. Are they so confident there will be fewer problems with future tests?

The Pragmatic View

Experts in educational policy may continue to debate the ways in which pressures for reform or accountability affect the educational process. Individuals in the trenches of local school systems seldom have debates like this. The explosion of state-mandated testing has left us with little time for discourse. Instead, the more pragmatic among us ask, "Can we stem the tide of new tests? If not, how can we make these tests less politicized? How can we reduce the negative and increase the positive impact of these tests? How do we manage the volume of testing? Are any of these tests useful to teachers?"

> By designing tests primarily to serve as tools of reform or accountability, policymakers miss opportunities to directly, rather than potentially, affect instruction. Use of fewer and shorter tests, samples of students, and tests at fewer grades would be adequate for accountability purposes. *If policymakers have no plans to directly affect instruction, they could certainly mandate fewer tests at a much lower cost to the public.*
>
> State-mandated census testing of every student may never provide excellent diagnostic information. When such testing is too time consuming, it drives out the chance for curriculum-based assessments. Sampling the smallest possible number of students needed for accountability is an under-utilized alternative to census testing.

Accountability pressures will continue for the foreseeable future, however. Our challenges are to understand the political pressures, to mediate some of these pressures, and to develop assessments that enhance rather than distort efforts to reform and improve instructional practice. This paper will also consider whether the goals of providing useful information to teachers and minimizing *noninstructional* testing (e.g., testing solely for reform or accountability) are incompatible with the purposes and politics of state testing plans. I will primarily focus on the testing program in North Carolina, because it provides an excellent example of the distortion political forces sometimes impose on assessment.

Politicians View Things Differently

Policymakers and educators hold different underlying beliefs about what is wrong with education and how to fix it. These underlying belief systems are a source of conflict that clearly affect policy decisions about the development and implementation of assessment systems, as well as the ways local educators react to mandated assessment programs.

Let's compare the two belief systems and how they conflict. The first is the political viewpoint. The rhetoric and actions of the North Carolina General Assembly regarding the state's testing program can serve as an example. The majority of the North Carolina legislature—and the majority of the state's public as well—seem to hold the following beliefs:

- Achievement is too low.
- Raising standards will increase student achievement.
- Ranking and comparing individual students, teachers, schools, and districts will increase student achievement.
- Teachers and local administrators are not very competent or motivated.

The political climate in many other states is similar. For example, Smith (1991) states that "...none could deny that the dominant public perception is that Arizona schools are failures, the teachers are not particularly hardworking (test scores 'prove' this), and the educational bureaucracy is inept."

Local educators generally do not share these perceptions. By contrast, these educators represent a second belief system that includes the following:

- External forces are relevant only when they change what teachers do; truly meaningful educational reform must occur at the classroom level.
- Policymakers have a compelling responsibility to provide teachers with test information that will enable them to improve instruction.
- Comparing and ranking students, teachers, schools, and districts often lead to increases in test scores without real increases in learning.
- Students spend too much time on testing for noninstructional purposes, including "pure" accountability.

These two belief systems have different implications concerning the best way to facilitate change. In the political arena, information about student performance leads to demands for change, and policymakers use humiliation, competition, and sanctions to inspire those who fall behind. In the educational arena, the importance of test data depends upon the extent to which it helps teachers improve their instruction.

Later in this paper, I will offer suggestions for reducing the political influences on testing while *maintaining* accountability and reform efforts. Progress will be difficult, however, until some of these differences in basic beliefs are reconciled.

Factors That Increase Political Influences

Let me offer some observations from the trenches—hypotheses, if you will—about factors that increase the political influences. Recognizing the factors may help us be more proactive in addressing the types of problems outlined in this paper. I believe political influences will increase when:

Governance structures are convoluted. For example, the diffusion of responsibility between the state board, state superintendent, and governor in North Carolina leaves a gap in leadership that is filled by the legislature. Partisan politics can further

complicate matters.

Student achievement is low, either in reality or in relation to other schools, districts, or states. Public pressure increases when test scores are perceived as low.

Leaders are elected or hired to "Clean up the mess." When the public perception is that the state, district, or school is a mess, the climate almost precludes a constructive, proactive approach to change.

Accountability is based on comparative models. Ranking schools or school systems adds to the public's scrutiny of test results, as well as misinterpretations of test data.

Large discrepancies between outcomes for different groups contribute to racial or class tensions. Large achievement differences between groups of students create a climate in which test scores are used inappropriately to document the "quality" of students and schools. Whereas some constituents may perceive the discrepancies as indicating inequities in the educational system, others may use the discrepancies to justify discrimination against low-scoring groups of students; for example, "those children are going to lower our scores if you reassign them to our school."

Most of these factors are present in the political arena in North Carolina. To illustrate some of the impact of these factors, let's first examine the educational governance in the state, and then some specific features of the state's testing programs.

Political Climate in N.C.

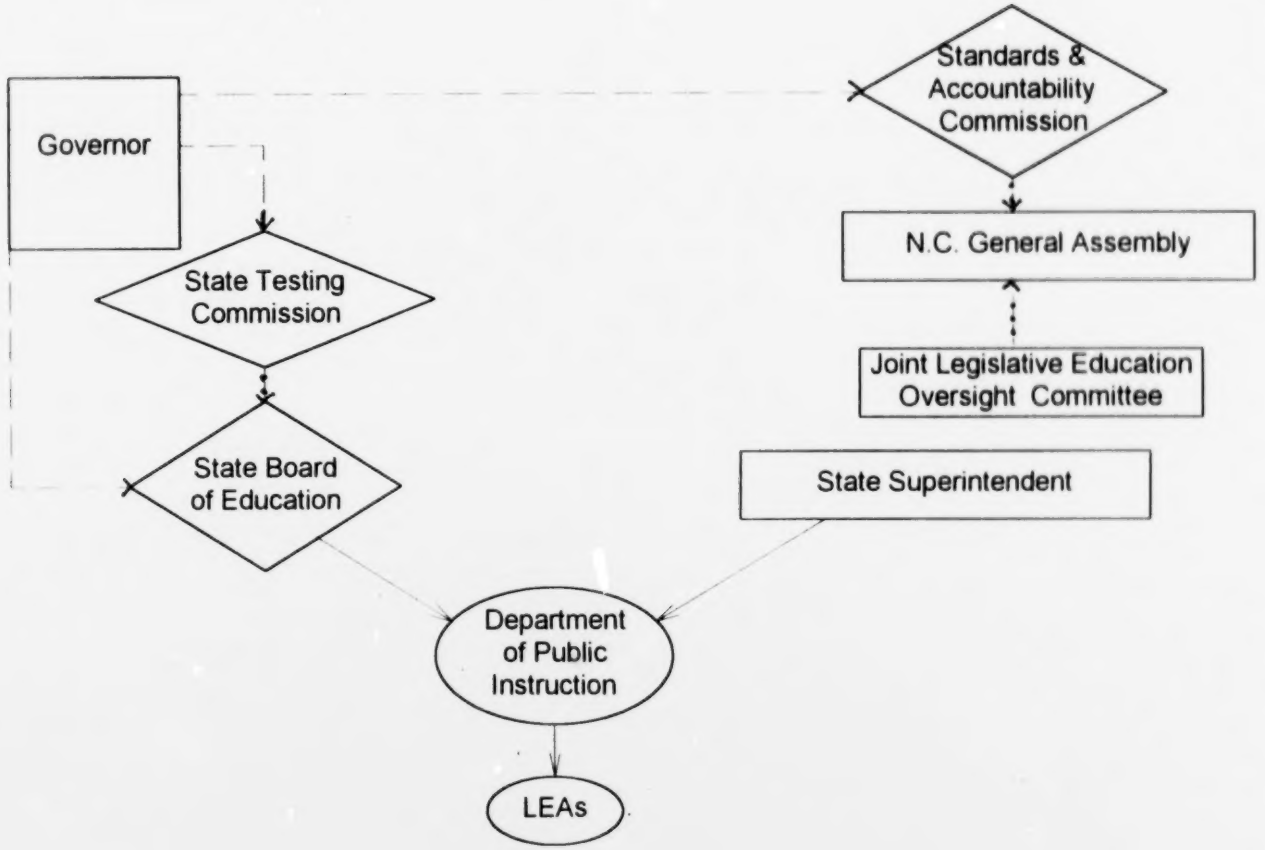
At best, educational governance in North Carolina is convoluted. At worst, it is a two-headed monster. Exhibit 1 shows some of the hierarchies involved in setting educational policy. Because the governor and state superintendent are both elected, they can belong to opposing political parties. Conflicts between the governor-appointed state board and elected state superintendent are not unusual. The governor also appoints the members of a State Testing Commission, to advise the state board on testing issues, and some members of the Standards and Accountability Commission, to make recommendations to the legislature.

In addition, the state superintendent may have ambitions for higher office and/or have little prior knowledge of education. As in many other states, gubernatorial and legislative candidates in North Carolina have made education a primary issue.

The political landscape is further complicated because achievement scores in North Carolina rank among the lowest of all states on many indicators scrutinized by the public. Scholastic Aptitude Test (SAT) scores for 1993 show the state ranking 48th, although most of the public is unaware the participation rate for North Carolina is higher than average. National

Exhibit 1

Educational Governance in North Carolina



Elected

Appointed

Advise

Appoint

Directs

April, 1994

Assessment of Educational Progress (NAEP) scores are similarly low.[2] In 1989, when the state legislature mandated the replacement of a ten-year-old, nationally norm-referenced test (NRT) with a state-developed, criterion-referenced test, many educators welcomed the chance for the state to develop tests to measure higher-order thinking skills and include "authentic" measures of what students can do. Many members of the public, however, saw this move as an attempt to hide how poorly North Carolina students scored in comparison to students in other states. Parents complained that, "No one cares how well their child did on a state test when that state is at the bottom."

The political environment in North Carolina affects the assessment program in various ways. In some cases, policymakers need data that make the state's achievement look desperately poor; in other cases, they need to show progress, albeit not too much. In this environment, with low SAT and NAEP rankings on the front page of the newspaper, a candidate or first-term state superintendent may initially emphasize the low achievement in North Carolina. Emphasis on low achievement not only prompts the flow of finite state financial resources to education, but also points out the need for strong leadership at the top. Later, during re-election campaigns, candidates must show that they have helped bring about progress, but there is more they need to do.

As part of the pressure to increase accountability in 1989, the legislature mandated state-developed report cards on each Local Education Agency (LEA) by 1990, followed by a more recent mandate for report cards on each school building in the state by 1996. The LEA report cards rely primarily on state test data; the building-level report cards produced in 1996 will do the same. Before these report cards, however, school systems were already compared on their performance on various tests, including optional tests like the SAT.

Assessment in N.C.: A Full Schedule

The political climate has resulted in a plethora of state-mandated tests in North Carolina. The state's assessment system is described in Exhibit 2. This system of tests may be one of the largest in terms of the number of state-mandated tests. For example, the state has decided to participate in the NAEP state-level testing, *and* to give a nationally-normed test (Iowa Tests of Basic Skills - ITBS) to a sample of students. A cynic might argue that this decision lets state officials quote North Carolina's poor NAEP rankings when it suits them, or its median ITBS percentiles around 50 when it suits them better. *In any case, the need to minimize noninstructional testing was less important than providing additional data to the state.*

---

[2]Ironically, the one measure on which North Carolina students appeared to achieve satisfactorily, with averages above the national norms, was the one measure the state legislature decided to replace.

Exhibit 2

State Testing Program in North Carolina
April, 1994

Test	Purpose	Grade Level(s)
N.C. End-of-Grade Test	To measure achievement in Reading, Mathematics, Science, and Social Studies	3-8
N.C. Writing Assessment	To measure achievement in Writing	4,6,8
Norm-Referenced Test (Iowa Tests of Basic Skills)	To permit comparisons of N.C. student achievement with national norms	5,8 Sample of students
National Assessment of Educational Progress (NAEP)	To contribute to national data on achievement (national sample) and to permit comparisons of N.C. student achievement with achievement in other states (state sample)	4, 8, 12 (national sample) 4 (state sample)
N.C. High School End-of-Course Exams	To measure proficiency in 10 high school courses; e.g., Chemistry, Biology, English I, English II, U.S. History	High School
N.C. Competency Test	To ensure students can perform basic math, writing, and reading tasks prior to high school graduation	10
Preliminary Scholastic Aptitude Test (PSAT)	To provide students with the experience of taking a test similar to the SAT and provide diagnostic information about areas for improvement	10
Scholastic Aptitude Test (SAT)	Although optional for students, LEAs are ranked and compared on this test "indicator"	11, 12 Optional
Aptitude Test (Test of Cognitive Skills, Otis-Lennon School Abilities Test, or other approved test)	To screen students for eligibility for the state's Academically Gifted program	2,5
Field Tests	To gather data on potential new items for EOG, EOG, and state item banks	3-12

Let's look at what happens to individual children. Eighth grade students in Wake County, North Carolina usually take *all* of the following state-mandated tests:

- N.C. End-of-Grade Tests in four subjects: Reading, Math, Science, and Social Studies (multiple choice and open-ended components);
- N.C. Writing Assessment; and
- N.C. End-of-Course Test in Algebra or Geometry.

Eighth grade students also usually take one of the following tests:

- Iowa Tests of Basic Skills (state sample); or
- National Assessment of Educational Progress (national sample); or
- Field tests for new state tests or state item banks (large state sample).

Clearly, North Carolina students take a large number of state-mandated tests. As Nolen, Haladyna, and Haas (1992) point out, measuring student achievement using multiple indicators is usually preferable to using a single indicator. Use of multiple indicators may be more resistant to "pollution" of scores through potentially questionable methods of raising scores, such as teaching only the tested curriculum. Multiple indicators can also yield a more complete picture of student achievement than any single indicator. Of course, use of multiple indicators is more important if the indicators are valid measures and they lead to improvements in instruction. *The large number of tests in North Carolina, however, has almost eliminated the opportunity for LEAs to add other, more informative assessments to a testing schedule that is overflowing, making it particularly important that the state's tests provide useful information. Arguably, the political pressure for accountability has resulted in an excessive amount of state-mandated testing in North Carolina.*

Unfortunately, according to the criterion of providing information that will improve teaching and learning, all but one of these tests for eighth grade students are **noninstructional** tests. The exception is the End-of-Course test, which provides information to teachers on students' mastery of instructional objectives, as well as other information. The other five tests provide schools with either *no* feedback or such global information that the results cannot be used to pinpoint weaknesses in instructional programs. State officials may believe that the other tests serve instruction in some way, but teachers would disagree with them about utility. Again, if policymakers have no plans to directly affect instruction, they could certainly mandate less census testing at a much lower cost to the public.

Let's focus on some of the tests in North Carolina, and look at how making different decisions during development and implementation could have enhanced the instructional use of the tests. As a pragmatist, I do not believe these mandated tests will disappear, because political pressures for accountability and reform are great. But the changes I will suggest would not have precluded fulfilling other purposes of the tests, including reform and accountability.

End-of-Grade (EOG) Testing

The most recent major new testing effort in North Carolina involves the End-of-Grade (EOG) tests. The EOG tests represent the state's best effort to date at exemplifying the new kinds and levels of achievement expected of students. Thus, state officials intended for the EOG tests to help reform education throughout the state. In addition, the tests were intended to measure and hold students, schools, and school districts accountable for educational performance.

The EOG tests consist of multiple choice and open-ended questions in reading, mathematics, science, and social studies. Students in grades 3-8 took the test for the first time in May, 1993. The EOG tests replaced the norm-referenced tests used for the previous decade. Students usually take the EOG tests in five testing sessions over a two-week period of time. Many educators considered the tests to be excessively long. In truth, the EOG tests took about the same amount of time as an NRT, but students found the EOG tests more grueling because of their difficulty.

Some good features of the new EOG tests deserve mention. First, EOG tests contain longer reading passages, because longer passages are considered more authentic and challenging (see Exhibit 3 for typical released items). Second, the EOG tests emphasize higher-order thinking skills, even on the multiple choice components of the tests. In addition, each student answers a sample of 10 open-ended items, with 30 open-ended items given at each grade across an entire class. (Thus, the EOG tests demonstrated that giving a sample of items to each student can increase the total item pool at lower cost in terms of student time, but with a corresponding loss of student-level information. Open-ended responses were not reported at the student or class level).

Finally, in December, 1993 schools received diskettes of scores and software that allowed them to disaggregate their scores by race, gender, and a variety of other factors. In spite of the delay in receiving the diskettes, schools throughout the state have used the diskettes extensively to identify groups of students who did not score as well as other groups.

Although these improvements over previous tests were noteworthy, local educators raised several concerns about the EOG tests. The concerns and limitations made the EOG results almost useless to the classroom teacher, and consequently diminished the impact of the test. The following sections describe some of these concerns.

Scheduling. State officials decreed that students must take the EOG tests during the last few weeks of school. This scheduling meant that results would not arrive in time for teachers to use them before the students left for the summer. Furthermore, the results would arrive too late to identify students needing summer remediation. Suggestions to move the testing window to earlier in the spring, however, fell on deaf ears. Instead, state officials contended, "The EOG test scores should reflect a specific year's performance, and you can't measure that in March or April." This decision on scheduling is an example of

Exhibit 3

SAMPLE RELEASED ITEMS
N.C. End of Grade Tests

Grade 3, Open Ended Item:

If you give a clerk $1.00 for a pencil that costs 76 cents, how much change should you get back? _____

Show or list 3 different sets of coins you could receive as your correct change from the clerk.

First way Second Way Third Way

Grade 3, Multiple Choice Item:

What is the sixth number in the pattern?

16, 26, __, __, 56, __

A 36
B 46
C 66
D 76

accountability—the question about how students performed at the *end* of the year—taking precedence over the needs of teachers and students.

Norm- vs Criterion-Referenced Tests. The new EOG tests were advertised to be criterion-referenced tests (CRTs) rather than norm-referenced. This change paralleled the philosophy that educators should first specify what students need to know and be able to do and then, second, assess them on those concepts and skills. Educators familiar with assessment, therefore, expressed surprise when they learned that results of the new test would include individual student percentile scores—based on state norms—for each student on this "criterion-referenced" test. This percentile method of reporting facilitates comparing and ranking of students, schools, and LEAs. Often, comparing and ranking are political, rather than educational processes.

Technical Issues. While only limited technical information is available, even this limited information has caused concern. The developmental scale scores on the tests have a range of 100 points; when the tests were scaled, the average scale score increase for a whole year of instruction was about five points for each of six grades, which seems rather narrow and may indicate an imprecision in measurement. The size of the standard error caused even greater alarm because it is generally as large as, or larger than, one year's growth.

Labels That Suggest Retention-In-Grade. The new EOG tests provide achievement level scores for each student, similar to the basic, proficient, and advanced levels used by NAEP. The EOG tests classify students into one of four achievement levels (see Exhibit 4) developed during field testing of the EOG tests. Level scores were based on teachers' predictions of students' future success. Although the achievement levels were an important component of the criterion-referenced nature of the EOG tests, two problems with the achievement levels caused concern.

First, labels for the four levels seem to suggest that students scoring at the two lower levels should be retained. Public reaction to statistics reporting the percentages of Level 1 and Level 2 students, therefore, was predictable. Newspaper editorials and letters exhorted schools to avoid social promotion and make students perform above an acceptable level before promoting them. Thus, despite what educators know about the harmful effects of retention-in-grade, the urge to slap an "accountable" label on students has led to pressures that could result in unwise educational decisions (Shepard & Smith, 1989).

Second, achievement level scores were not validated by determining how well the field-tested students actually performed at the next grade level. This type of validation would have taken additional time. Scale score data would not have been as powerful as level scores for communicating with the public. While delay would have been technically preferable, the politics of a major state testing effort such as this one apparently precluded taking the time to do things right. This shortcut seems misguided, given the intense focus on Level 1 and Level 2 students previously described.

Exhibit 4

DESCRIPTION OF ACHIEVEMENT LEVELS

N.C. End of Grade Tests

Level 1. Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.

Level 2. Students performing at this leveel demonstrate inconsistent mastery of knowledge and skills in this subject area and are minimally prepared to be successful at the next grade level.

Level 3. Students performing at this level consistently demonstrate mastery of grade level subject matter and skills and are well prepared for the next grade level.

Level 4. Students performing at this level consistently perform in a superior manner and are clearly beyond that required to be proficient at grade level work.

Publishing Results by Class. In 1993, parents received a report indicating the performance of their child, their child's school, the school system, and the state (see Exhibit 5). In 1994, the state intended to add EOG classroom-level averages on reports sent home to all parents. The purpose of reporting scores for each teacher would be to hold teachers accountable for their students' performances.

State assessment officials have agreed that including results for each classroom teacher can be unfair, particularly when students are not randomly assigned to classes. Psychometricians who worked on the test privately agreed that this reporting of scores on very small groups of students also might be technically questionable because the standard error of measurement on the test is bigger than the average scale score increase for a whole year of instruction. In the meantime, policymakers who have pressed to include class results seem unconcerned that our *best* teachers often teach a disproportionate number of difficult-to-teach students and that low scores from such classes might reflect better-than-usual results for those students. *Rather, political forces appear to be driving the decision to include this class-level information so that teachers will be "held accountable."*

Lack of Instructional Information. Another concern was the paucity of instructional information produced by the EOG tests. The class summary report for teachers indicated only how well students did on various broad goals (see Exhibit 6). This brief report was generated for each class and grade level. The report contained no information for detailed objectives. For example, one of the test objectives is to "apply, extend, and expand on information and concepts." Teachers and administrators have indicated that knowing their students do poorly on this objective does not help them improve instruction because the objective is too broad. As a teacher, how would you know what parts of that objective your students had trouble with? How would you change your instruction to "fix" the problem?

State testing officials indicated that more test items would be needed to allow more specific reporting from EOG tests, even at the classroom or building level. Yet, in reading alone, students in each class took between 168 and 198 multiple choice items across three forms of the test. *This number of items should have been sufficient to provide more detailed information, at least at the school level, if improving instruction had been considered as a primary goal when the test was first developed.*

Later, state curriculum officials reportedly acknowledged that the match between EOG test items and more specific objectives in the state curriculum had never been completed. The failure to complete this matching process is inconsistent with the reformers' philosophy of *first* determining what students should be able to do and *then* designing measurements that match. But, micromanaged mandates from the legislature specify that new tests will be given by a certain date, and nothing like a lack of time to match test items to the curriculum can stand in the way.

While the EOG tests will never replace classroom diagnostic testing completely, providing

Exhibit 5

End of Grade Testing
N.C. Public Schools
Parent/Teacher Report

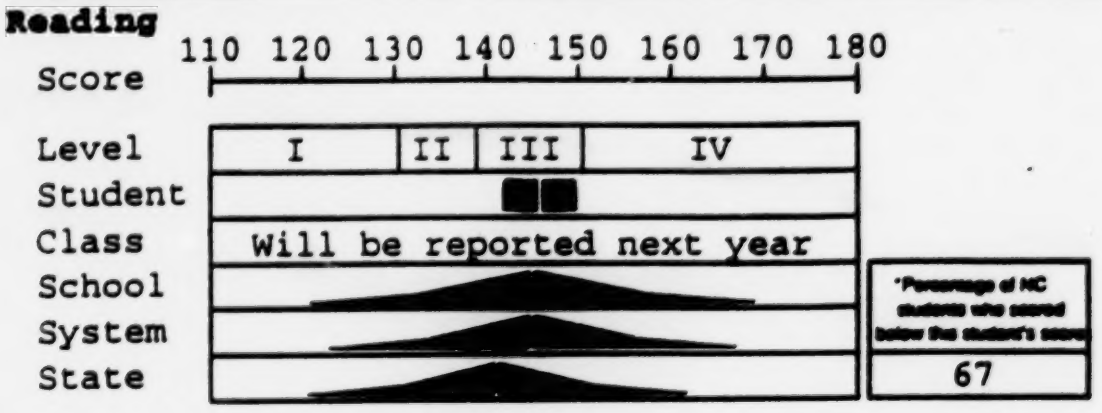


Exhibit 6

School or Class Report Format
N.C. End of Grade Tests

END-OF-GRADE SUMMARY GO GRADE 3 — END-OF-GRADE SUMMARY GOAL REPORT 1993:

LEA: 920

Printed by: SchoolName = A ELEMENTARY

	Score Mean	Number of Observations	Items/ Score	Percent correct
Reading	40.7	97	56	72.7
All Reading Items	122.0		168	72.6
GOAL 1: Use strategies and processes that enhance control of communications skills development			0	.
GOAL 2: Use language for the acquisition, interpretation, and application of information			130	75.9
OBJ 2.1: Identify, collect or select information and ideas			58	80.8
OBJ 2.2: Analyze, synthesize, and organize information and ideas and discover related ideas, concepts or generalizations			48	72.3
OBJ 2.3: Apply, extend, and expand on information and concepts			24	71.4
GOAL 3: Use language for critical analysis and evaluation			38	61.4
Mathematics	62.2	97	80	77.8
All Mathematics Items	186.5		240	77.7
Math Computation			36	91.1
Math Applications			204	75.4
Goal 1: Identify and use numbers to 1000 and beyond			24	78.0
GOAL 2: Understanding and use of geometry			24	77.3
GOAL 3: Understanding of classification, pattern, and seriation			24	73.1
GOAL 4: Understand and use standard units of metric and customary measure			36	76.5
GOAL 5: Use mathematical reasoning and solve problems			36	68.1
GOAL 6: Demonstrate an understanding of data collection, display, and interpretation			24	78.0
GOAL 7: Compute with whole numbers			72	84.6

	A	B	C
NUMBER OF STUDENTS TAKING FORM	31	34	32

simple information on a legislatively mandated schedule apparently took precedence over reporting information in ways that would provide at least some useful information to teachers.

Hindering Reform. New approaches to instruction and curriculum require new approaches to assessment, unless we want teachers to continue teaching to the old tests. The new multiple choice tests in science and social studies were not ready in time for the spring, 1993 EOG testing. Instead, the new tests incorporated the old, out-of-date multiple choice tests in science and social studies as part of the "new" tests. The old test questions no longer match newer approaches to instruction and curriculum, and therefore the science and social studies tests produced useless results for that year, from an instructional point of view. The decision to administer and report scores from these old tests was apparently made because science and social studies scores are part of the state's "accountability" program. Yet, testing the previous year took place in March, and therefore new May test scores were not comparable anyway. We can *hope* that teachers taught with new methods and curriculum, but it is likely that many taught the same old way because that was what was tested. *In addition, there is no way to justify the lost instructional time for administering these tests, or the misrepresentation to the public inherent in reporting the scores. The politics of accountability collided with reform efforts, in this case.*

Summary. In trying to serve multiple purposes, the North Carolina End-of-Grade tests have demonstrated several limitations. These tests do a better job of exemplifying the goals we have for students than did previous state tests, but at some sacrifice of instructional utility, technical merit, and fairness to teachers. Legislatively imposed deadlines may have hindered state education department officials improving the tests, but some changes are still possible.

Among the desirable changes for North Carolina's End of Grade Tests would be:

- testing in the fall or earlier in the spring,
- omitting percentiles and classroom-level results on parent reports,
- validating and revising the nomenclature for the achievement levels, and
- increasing the amount of instructional information provided to teachers.

These changes would enhance the utility of the tests for instructional purposes, and would not negatively impact other purposes of the test. If these changes cannot be implemented, state policymakers should consider sampling as a way to reduce the costs in terms of student time and money.

Writing Assessment

The North Carolina Writing Assessment is another state-mandated test in which modifications could enhance the usefulness of the test to teachers. North Carolina students in grades 4, 6, and 8 take the North Carolina Writing Assessment annually. The test consists of a single

writing prompt at each grade level, to which students respond in a single testing session. The tests measure one of the five types of writing: narrative, descriptive, persuasive, clarification, and point-of-view writing. The writing tests are scored holistically and are reported at the student, class, building, and district levels. Many educators support the "performance" nature of this writing test, but many also have concerns about specific features of the writing assessment.

Reliability and Reporting. Single-prompt, open-ended assessments such as this writing assessment are estimated to have the reliability of a four-item multiple choice test. Normally, this reliability is too low for reporting student-level scores, but the state has chosen to report scores for individual students on the N.C. Writing Assessment, implying that the test is a valid measure of how well a particular student can write. Parents are sometimes alarmed unnecessarily because their child's score report reflects a poor performance on a particular prompt. State officials caution that the results should not be used at the individual student level, yet they continue to report them at that level, presumably to make sure students and teachers feel accountable. *If* giving a test that is unreliable at the individual student level is important, then I propose not reporting scores at the individual student level, but only at the classroom level and above. *As an alternative, recognizing that the North Carolina Writing Assessment is inappropriate for student-level measurement, a sample of students would be sufficient for accountability purposes.*

Scoring. The holistic scoring method the state employs for scoring the writing tests yields little useful information for teachers. Essentially an accountability tool, holistic scoring does not distinguish between low scores due to vague vocabulary, poor organization, weak arguments, or misunderstanding the topic. *Multi-factor, analytic scoring would give teachers information that might actually improve their instruction. Holistic scoring, of course, works just fine for comparing and ranking LEA's and schools.*

Restrictions on the Writing Process. Finally, this test reflects the impending collision between accountability and reform efforts. Specifically, the administrative procedures for the writing assessment do not exemplify what we want students to be able to do; i.e., reflect, research, outline, draft, review, and revise their writing over a period of time. Why not allow students to use reference works or revise their papers in a second testing session? Even though their papers may not change substantially, the test would be more authentic. Eventually, students who use a computer for most writing projects will need to be tested using a computer. *Accountability might suffer by making the task more authentic, but not as significantly as instructional reform suffers under the current testing procedures.*

High School End-of-Course Tests

The North Carolina End-of-Course (EOC) tests comprise another component of the state's testing program. The state mandates that students take EOC tests in 10 high school subject areas, including a few high school courses taught in middle schools. Half of the 10 courses with mandated EOC tests are required courses for all North Carolina students, while the

remaining half are optional courses. Most of the EOC tests are multiple choice; the exceptions are the English II essay test, an open-ended portion of the Algebra I test, and the "proof" portion of the Geometry test.

Expanded Reporting of Scores. In an effort to hold individual students more accountable, state policymakers are currently discussing plans to require inclusion of all EOC test scores on students' high school transcripts. Colleges and employers are expected to use the scores in hiring and selection decisions, although the tests have not been validated for this purpose. In addition, state policymakers have discussed establishing minimum passing scores on each EOC test, and using a battery of EOC tests to replace the current minimum competency test for graduation. If these proposed changes are implemented, some students may fail to graduate because they fail an EOC test. This raises issues of state versus local control over who graduates.

Mandated Grading Policies. State officials are also discussing a proposal to increase the pressure on low ranking LEAs by requiring the EOC test scores to count as a fixed percentage (one-third) of students' final course grades and by setting cut-off test scores for each letter grade. (Currently, the percentage the EOG tests count and the letter grade cutoffs and distributions are left up to LEAs). If these changes are implemented as planned, the state will essentially have established grading policies for LEAs. In many low ranking school systems, these changes would raise the cutoffs for As and Bs . In many systems that are failing large numbers of students, the changes would lower the cutoff for Ds (versus Fs).

Effect on Instruction. Ironically, these proposed changes call into question why state cutoffs for passing and graduation are needed that are lower than cutoffs currently used by many LEAs for passing the courses. The effects will differ across LEAs. For example, in Wake County, which is usually one of the highest ranked systems, the proposed change will actually increase the number of As in most courses, thereby lowering the standards for the course. In spite of these easier grading standards, the changes will probably water down Wake County's more rigorous curriculum because teachers will feel they should spend more time on "just" the state curriculum when the test will be weighted so heavily.

Clearly, something has gone awry in North Carolina. The same policymakers who are urging a more rapid shift to site-based management plan to intrude extensively into teachers' grading policies while substantially increasing the stakes of the testing in other ways, including ways that are technically questionable.

What Can Be Done?

In their discussion of using research in high stakes public policy environments, Archer et al. (1992) discuss ways to reduce conflicts between policy researchers and policy makers. While the authors suggest that conflicts are inevitable because of the adversarial culture of regulatory policy makers, they do include at least two suggestions that are applicable here:

- Use formative rather than summative evaluation approaches; and
- Use a prospective rather than retrospective focus.

In a high stakes environment, the effect of asking "how many achievement score points did students earn"—essentially a summative approach—only serves to increase the tendency to blame someone for the scores or to credit someone with the success. A formative approach, that of asking "what are the areas of comparative weakness, or what do kids need to know that they don't know" would refocus the question and lower the thermostat a bit. For example, fall testing would reduce the stakes, in comparison with testing at the end of the year. I would submit that, in spite of all the current rhetoric about the need to redefine what it is that kids need to know, that we as a nation spend more of our assessment dollars on questions of "how high?" than we do on questions of "what?"

I will not offer any practical suggestions for how to address issues of educational governance or differences in underlying beliefs, because people employed by the organizations involved may not have the freedom to work on these issues. I will, however, offer some suggestions that may make a difference in other areas. The suggestions include some areas where we can control our own actions, as well as areas where we may be able to influence actions of others. Educators in local districts can take several approaches:

- **Maintain your own professional integrity.** The superintendent in Wake County was asked a few years ago whether the district should continue buying the old edition of consumable test booklets for the last year of norm-referenced testing or buy the newer test edition and risk having scores decrease because of the tougher questions and newer norms. There were no cost implications because all materials were consumable. His only question was, "Which test would give better information to teachers?" When told the newer test was better, he said, "Fine, buy it. We'll deal with the press questions if the scores go down, but we need to administer the test that is best for instruction." That's leadership. It involves risk, but if we all take a few more risks, things <u>will</u> improve.

- **Organize to voice concerns.** The testing directors of several large districts in North Carolina meet a few times each year to discuss concerns about the testing program. These LEAs represent a large proportion of the total student enrollment in the state. Political forces often preclude state officials from agreeing to some requests the group has made. However, we recently learned that the state has agreed to omit class averages from parent reports. We have also shared ideas with other LEAs about ways to implement state mandates with as little harm as possible.

- **Make noise.** LEAs can point out when politics are being placed before educational concerns. Call or write state legislators, and tell concerned parents to do the same. Use the media. Use the PTAs and teachers' unions or bargaining groups. Document the amount of time and money spent on noninstructional tests. Raise technical concerns in understandable ways. Point out costs associated with census testing.

- **Remember that "those who live by the test scores, die by...."** De-emphasize test scores and encourage your superintendent to do the same. If you take a "ho, hum, the scores went up because the writing assessment prompt was easy" attitude, it is much easier to say, "ho, hum, the scores went down again."

State officials can also take some steps to improve things. They can include more teachers and other LEA staff in developing report formats and test frameworks, rather than just as "item writers." The more input state officials have from teachers, the more likely they are to produce tests and reports that are at least somewhat likely to make an instructional difference.

Also, if we acknowledge that teachers are going to teach to the tests, state officials can ensure each test exemplifies what students should be able to do. In North Carolina, the new EOG tests in reading and math accomplished this goal, while the decision to use old science and social studies tests was potentially damaging.

Clearly, the purposes of local assessments—to improve instruction—may differ from the purposes of externally mandated assessments. Externally mandated assessments may not only serve as tools of accountability and reform, but they may also be part of agendas to make achievement look bad, good, or somewhere in between. They may affect elections, public opinion, and efforts to reform education.

In my vision for the future of state and national assessments, decisions will be based first and most importantly on what is best for students and teachers. This reform will mean sampling, scoring, and reporting in ways that are different from current methods. The process of developing and implementing assessments will reflect the awareness that improving instruction ultimately depends on the classroom teacher. The majority of the time and resources devoted to assessment will focus on providing information useful for instruction.

We must continue to work to reduce the amount of time spent on noninstructional tests, but simultaneously work to make noninstructional tests more useful and less harmful because these tests will be with us for the foreseeable future. For our educational system to thrive, all constituencies involved in assessment must work cooperatively to develop optimal assessment systems. If these new assessments are not valid, reliable, useful, cost-effective, and consistent with new directions in curriculum and instruction, their shortcomings will eventually undermine the assessments, and in turn further erode the public's confidence in education. More thoughtful decisions by policymakers can help improve our current system, even within the political context. It is not just desirable that we refine our current assessment systems, it is imperative.

Appendix 1

Do Accountability Efforts Help or Hurt Instruction?

There is a growing body of research that suggests high-stakes accountability programs hurt instruction. Lengthy discussions of this issue are beyond the scope of this paper but three recent reports are worth mentioning. Corbett and Wilson (1991) discuss "reform by comparison" of test scores, in which ranked comparisons of districts' test scores stimulate action at the local level, particularly for low-ranking districts. Under this approach, the authors conclude, the type of action taken to raise scores is less important than taking some action. The authors discovered that the majority of educators they interviewed felt that their districts' actions to increase achievement scores were inconsistent with their goal of educational improvement, which was the original purpose of the reform. Attempts to raise scores through means *other than* instructional improvement were reported as more common responses to accountability pressures. Nolen, Haladyna, and Haas (1992) recount similar conclusions in their study of the impact of high stakes testing in Arizona.

Loofbourrow (1993) found that teachers in one California school she studied re-aligned their curriculum to focus on preparing students for the state Writing Assessment. The test was intended to encourage writing across the curriculum, but English teachers at this school were assigned the responsibility of preparing students for the test. Why? Because schools with high scores got extra funding, and probably also because the writing prompts did not really reflect the interdisciplinary approach that was advocated by policymakers who mandated the test.

When the Public School Forum studied reform efforts in six southeastern states, the researchers were interested in the impact of accountability mandates on schools and systems (Dornan, 1993). Their research questions included "Has accountability changed the way practitioners approach their jobs?" and "Are consequences like probation and takeovers making a difference?" The study focused on a total of seven low-performing districts. The researchers found that these systems were not able to meet new, higher standards, which was not surprising because they had trouble meeting existing, easier ones. In addition, state efforts were not geared towards helping schools like these succeed. Instead, the states' focus was on "keeping score". Assistance was provided from the states only when state takeover was becoming imminent, and when the LEA improved performance, the assistance was withdrawn. Finally, the social needs of these schools made the curriculum and staff development assistance the states offered seem woefully inadequate. For example, one school has 97% of its students qualifying for free lunch. The researchers concluded that true change in the lives of these students would require a joint effort between educational, social, law enforcement, and development agencies.

References

Archer, D., Pettigrew, T.F., & Aronson, E. (1992). Making research apply: High stakes public policy in a regulatory environment. American Psychologist, 47(10), 1233-1236.

Airasian, P.W. (1993). Policy-driven assessment or assessment-driven policy? Measurement and Evaluation in Counseling and Development, 26(1), 22-30.

Allington, R.L. & McGill-Franzen, A. (1992). Does high-stakes testing improve school effectiveness? ERS Spectrum, 10(2), 3-11.

Bond, L., Friedman, L., & van der Ploeg, A., (1993). Surveying the Landscape of State Educational Assessment Programs". Washington, D.C.: Council for Educational Development and Research.

Corbett, H.D., & Wilson, B.L. (1991). Testing, Reform and Rebellion. Norwood: Ablex.

Cuban, L. (1993). The Lure of Curricular Reform and It's Pitiful History. Phi Delta Kappan, 75(2), 182-185.

Dornan, J., (1993). Rethinking accountability. The Forum Report, 8(2).

Hansen, J.B. (1993). Is educational reform through mandated accountability an oxymoron? Measurement and Evaluation in Counseling and Development, 26(1), 11-21.

Herman, J.L. & Golan, S. (1993). The effects of standardized testing on teaching and schools. Educational Measurement: Issues and Practice, 12(4), 20-25, 41.

Hymes, D.L., Chafin, A.E., and Gonder, P. (1991). AASA Critical Issues Report: The Changing Face of Testing and Assessment. Arlington, Va.: America Association of School Administrators.

Loofbourrow, P.T. (1993). Composition in the Context of the CAP: A Case Study of the Interplay Between Assessment and School Life. Berkeley, CA.: National Center for the Study of Writing and Literacy.

Madaus, G.F. & Kelleghan, T. (1993). Testing as a mechanism of public policy: A brief history and description. Measurement and Evaluation In Counseling and Development, 26(1), 6-10.

Mazzoni, T.L. (1993). The changing politics of state education policy making: A 20-year Minnesota perspective. Educational Evaluation and Policy Analysis, 15(4), 357-379.

National Educational Goals Panel, Goals 3 and 4 Technical Planning Group (1993). Promises to Keep: Creating High Standards for American Students. NEGP, 94-01, D-58.

Nolen, S.B., Haladyna, T.M., & Haas, N.S. (1992). Uses and abuses of achievement test scores. Educational Measurement: Issues and Practice, 11(2), 9-15.

Pommerich, M.; Billeaud, K., Williams, V.S.L. & Thissen, D. (1993). User's Guide for the North Carolina End of Grade Tests. Chapel Hill: Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

Popham, W.J. (1993). Measurement-driven instruction as a "quick-fix" reform strategy. Measurement and Evaluation In Counseling and Development, 26(1), 31-34.

Rudman, H.C. (1993). National testing or political testing: Is there a difference? Educational Measurement: Issues and Practice, 12(3), 5-10.

Shepard, L. & Smith, M.L. (1989). Flunking Grades: Research and Policies on Retention. New York: Falmer Press.

Smith, M.L. (1991). Put to the test: The effects of external testing on teachers. Educational Researcher, 20(5), 8-11.

Stiggins, R.J. (1993). Two disciplines of educational assessment. Measurement and Evaulation in Counseling and Development, 26, 59-63.

Williams, P.L. & Slawski, E.J. (1993). Educational policy and assessment standards: Can they ever meet? Measurement and Evaluation in Counseling and Development, 26(1), 59-63.